ISSN 0976-0962

**International Journal of Computational Linguistics and Applications**

**Vol. 6** **No. 1** **Jan-Jun 2015**

**CONTENTS**

*International Journal of Computational Linguistics and Applications*, Vol. 6, No. 1, 2015, pp. 25–44
Received 28/02/2015, Accepted 18/05/2015, Final 12/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

# On the Influence of Text Complexity on Discourse-Level Choices

ELNAZ DAVOODI
LEILA KOSSEIM
*Concordia University, Montreal, Canada*

ABSTRACT

*Text complexity can be reduced by making different choices at the lexical and grammatical levels. However, discourse-level choices may also affect a text's complexity. In a coherent text, explicit discourse relations (e.g.* CAUSE, CONDITION*) are expressed using discourse markers (e.g.* since, because*, etc.) that may be preferred for texts at different readability levels. In this paper, we investigate the differences in discourse properties of texts across readability levels. In particular, we investigate the effect of readability level on (1) the usage of discourse relations, (2) the usage of discourse markers and (3) the distribution of discourse markers signaling explicit discourse relations. Our analysis of the Simple English Wikipedia corpus shows that complex and simple texts seem to have the same distribution of discourse relations; however, these relations are expressed using different discourse markers depending on the readability level of the text.*

## 1. INTRODUCTION

In well-written texts, utterances are connected to each other using Discourse Relations (DRs) which allow the reader to understand the communicative intention of the writer. DRs (e.g. CAUSE, CONDITION) can be expressed either implicitly or explicitly. *Implicit* relations are not signalled using lexical cues such as *but*, *since*, *because*, etc. and must be inferred by the readers. On the other hand, *explicit* relations are signalled using specific terms

called Discourse Markers (DM). According to [21], DMs constitute strong clues to detect explicit relations, hence discourse parsers have used them as valuable features in order to identify DRs automatically (e.g. [13, 26, 10, 16]).

A text's discourse-level properties have been shown to be correlated to various dimensions such as their genre, their level of formality, their level of readability, etc. For example, Webber [27] and Bachand et al. [1] showed that the textual genre influences the choice of DRs. In order to produce texts at various readability levels, several techniques have been proposed to simplify texts at the lexical level (e.g. [7, 30]), the syntactic level (e.g. [3, 24]) and the discourse level (e.g. [25]). In particular, Williams [28] used "simpler" DMs to generate more readable texts for people with a lower level of literacy. In the process of text simplification, the writer's goal is to reformulate a text to make it easier to read and understand; however, its informational content should be preserved. Based on this assumption, we suspected that the simplification process should not change the semantic or logical relations between textual units; however, because DRs can be used for rhetorical purposes [17], the distribution of DRs may be different across text complexity.

One can view texts at different readability levels as translations of their "regular" counterpart. Using this perspective, we can argue that during the translation, translators may choose to use DRs and DMs differently in the translated text by adding or removing them or making implicit relations explicit or vice versa; all the while, preserving the meaning of the original text. For example, in the context of machine translation, Meyer and Webber [18] have shown that fewer DMs were used in the German or French translations of the Newstest 2012 parallel corpus[1] compared to its English counterpart.

In this article, we investigate the inuence of the readability level on the usage of explicit DRs and DMs. We have used the Simple English Wikipedia corpus [4] which has been used widely in text simplification and the related task of text

[1] http://www.statmt.org/wmt12/

compression (e.g. [29, 30]). The usage of explicit DRs and DMs as well as the distribution of DMs used as cues of such relations are analyzed. We used the log-likelihood ratio to rank the DRs and DMs with texts at various levels of readability.

This paper is organized as follows: Section 2 describes the corpus preparation to extract DRs and DMs. Section 3 presents the results of each experiment. Related work and discussions are presented in Section 4 and finally Section 5 presents our conclusions and future work.

## 2. CORPUS PREPARATION

To investigate the infuence of the readability level on the usage of DRs and DMs, these were extracted automatically from parallel corpora accross different readability levels.

### 2.1. *The simple English Wikipedia corpus*

Because they are manually annotated with DRs, the RST-DT corpus [2] and the Penn Discourse TreeBank (PDTB) [22] constitute two of the most widely used corpora for discourse analysis. However, these corpora could not be used in our work because we needed a parallel corpus across different readability levels. Instead, we used the Simple English Wikipedia corpus [4] which is a parallel corpus containing regular and simplified versions of Wikipedia articles. The simplified versions of the Wikipedia articles are meant to be more accessible to beginners learning English, such as students, children, adults with learning difficulties and people who are trying to learn English. These articles are typically shorter than their regular counterparts, and use simpler words and syntactic structures. The simplified articles were created by using their regular counterparts as a basis and following a set of simplification guidelines.[2] In particular, word choices are limited to Basic English[3], a 850-word auxiliary

[2] https://simple.wikipedia.org/wiki/Wikipedia: How to write Simple English page

[3] https://simple.wikipedia.org/wiki/Wikipedia:Basic English ordered wordlist

international language, and the VOA Special English Word Book,[4] a list of 1580 words. The guidelines are not only limited to lexical choices, but also suggest the use of simpler syntactic structures; such as avoiding compound sentences containing embedded conjunctive clauses.

The Simple English Wikipedia corpus was first created from Simple Wikipedia articles[5] in 2010. The first version of this corpus contains 137K aligned sentences pairs created from Wikipeda pages downloaded in May 2010. The latest version, released in 2011, contains two parts: a sentence-aligned part containing 167K aligned sentence pairs and 60K aligned articles. In our work, we used the aligned sentences of the latest version of this corpus.

### 2.2. *Labeling the corpus*

Because the Simple English Wikipedia corpus is not discourse-annotated, to label DRs and identify DMs signalling explicit DRs, we have automatically parsed the parallel sentences using the End-To-End PDTB-based discourse parser [16].

Several other publicly available discourse parsers could have been used (eg. [13, 10, 9]). We chose the End-to-End parser because we needed local discourse-level information that include the type of discourse relations (i.e. *implicit* or *explicit*), the name of the discourelation and the discourse marker when applicable. When the work was performed, the End-to-End parser was the best performing parser providing all these features. Although the parser can identify both explicit and implicit DRs, we only considered explicit DRs as the accuracy of the parser in detecting explicit relations is about 81.19% whereas for implicit relations the accuracy drops significantly. In addition, because we are interested in the usage of discourse markers which signal explicit DRs, implicit relations were not considered.

---

[4] https://simple.wikipedia.org/wiki/Wikipedia:VOA Special English Word Book

[5] www.simple.wikipedia.org

The End-to-End parser [16] uses the PDTB inventory of relations [22] organized into 3 levels of granularity. Level 1 includes four relations: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. In our experiment, we used the 2nd level that defines 16 relations, but only 12 relations were present in the corpus. In addition, the End-to-End parser uses an inventory of 100 DMs, but only 72 were actually present in the Simple English Wikipedia corpus.

Table 1 provides statistics about the annotation of the regular and simple versions of the Simple English Wikipedia corpus with the End-to-End parser. As shown in Table 1, the regular version of the sentence-aligned part of the corpus contains 167K sentences; however in the simple version, the number of sentences increases to 189K sentences. In the simple version, sentences tend to be shorter (18.45 words versus 23.36) and fewer DMs are used. In addition, the ratio of DM per token is tend to be lower in the simple version compared to the regular version (0.093 vs 0.098).

Table 1. *Statistics of the Simple English Wikipedia corpus*

| | **Regular version** | **Simple version** |
|---|---|---|
| # of sentences | 167,690 | 189,572 |
| # of DMs | 52,648 | 48,412 |
| token/sentence ratio | 23.36 | 18045 |
| DM/token ratio | 0.098 | 0.093 |
| DM/sentence ratio | 0.31 | 0.25 |

## 3. ANALYSIS

Once the Simple English Wikipedia was tagged with DMs and DRs, we analysed: (1) the usage of DRs, (2) the usage of DMs and (3) the distribution of DMs over DRs across readability levels.

### 3.1. *Effect of text complexity on the usage of DRs*

Once the parallel corpus was parsed with End-to-End parser [16], we extracted the explicit DRs in both the regular and the simple versions. In order to eliminate the effect of corpus size, we

considered the relative frequencies of DRs, then we performed frequency profiling using the log-likelihood ratio [23]. This measure allows us to compare the frequency of DRs across the regular and the simple versions and sort them according to the importance of their relative frequencies. The log-likelihood ratios themselves only provide a measure of which DRs are statistically more informative. The results are shown in Table 2 in decreasing order of log-likelihood ratio. The relations at the top of the table are therefore more indicative of the regular version, as compared to the simple versions of the corpus.

According to Table 2, the most differences stem from the relations of CONTRAST, CAUSE and CONCESSION; however in both the regular and the simple versions, the three most frequent DRs are CONJUNCTION, CONTRAST and ASYNCHRONOUS.

In order to verify if these changes are statistically significant, we first performed a normality test using the IBM SPSS software[6] to investigate the characteristics of our data set. According to this test, the relative frequency of DRs in the regular and simple versions is not normally distributed. Consequently, we have used the Wilcoxon test of statistical significance to see if the difference across the two corpora are statistically significant. The Wilcoxon test is a non-parametric statistical hypothesis test which is an alternative to the Student's t-test when the population is not normally distributed. According to this test, the differences in the relative frequencies of DRs are not statistically significant. As a result, we can conclude that the usage of DRs seems to be preserved across different readability levels of this parallel corpus.

[6] http://www-01.ibm.com/software/analytics/spss/

Table 2. *Relative frequency of DRs across regular and simple versions of the Simple English Wikipedia corpus sorted by log-likelihood ratio*

| Discourse Relation | Regular Version | Simple Version | LL Ratio |
|---|---|---|---|
| CONTRAST | 18.10% | 16.29% | 20.76 |
| CAUSE | 7.62% | 8.64% | 13.82 |
| CONCESSION | 2.88% | 2.33% | 12.49 |
| RESTATEMENT | 0.31% | 0.20% | 4.85 |
| CONDITION | 4.06% | 4.46% | 4.01 |
| ASYCHRONOUS | 14.76% | 15.31% | 2.22 |
| SYNCHRONY | 12.51% | 12.75% | 0.48 |
| EXCEPTION | 0.04% | 0.05% | 0.22 |
| LIST | 0.01% | 0.02% | 0.17 |
| CONJUNCTION | 36.52% | 36.72% | 0.12 |
| ALTERNATIVE | 1.75% | 1.78% | 0.06 |
| INSTANTIATION | 1.38% | 1.39% | 0.00 |

3.2. *Effect of text complexity on the usage of DMs*

Given that the usage of DRs seems to be preserved, we next turned to how they are signalled across readability levels. DMs can signal more than one DRs. For example, *although* can signal both a CONCESSION and a CONTRAST relation. In this experiment, we were interested in investigating the distribution of DMs over DRs.

Once all the DMs and DRs were extracted using the End-to-End parser [16], we constructed DM/DR pairs in order to disambiguate DMs that can signal more than one DR. As a result, we created a set of 119 unique DM/DR pairs. Then, we again used log-likelihood ratios to sort the pairs. Hence, a DM/DR pair with a higher log-likelihood ratio is more indicative of the regular version, as compared to the simple version of the corpus. Table 3 shows the 10 most discriminating pairs across the regular and simple versions.

Using all DM/DR pairs extracted automatically, we have again performed a statistical significance test in order to determine if the difference in the relative frequency of DM/DR pairs across corpora is statistically significant. Similarly to the first analysis (see Section 3.1), we first performed a normality test using the IBM SPSS software. The results revealed that

DM/DR pairs are not normally distributed across corpora. The relative frequency of some pairs such as *because*/CAUSE, *so*/CAUSE and *but*/CONTRAST is higher in the simple version, while it is lower for other pairs such as *thus*/CAUSE, *although*/CONTRAST and *while*/CONTRAST. The Wilcoxon statistical significance test showed that the relative frequency of DM/DR pair across different readability levels is statistically different. More precisely, the Wilcoxon test revealed that in the simple version of the Simple English Wikipedia, DMs are used less frequently than in its regular counterpart. This is an interesting finding as it seems to indicate that to make a text more accessible, the use of discourse markers should be reduced; hence not indicating discourse relations explicitly.

Table 3. *Relative frequency of DM/DR pairs across regular and simple versions of the Simple English Wikipedia corpus sorted by log-likelihood ratio*

| Discourse Relation | Regular Version | Simple Version | LL Ratio |
|---|---|---|---|
| because/CAUSE | 0.0280% | 0.0470% | 76.99 |
| thus/CAUSE | 0.0166% | 0.0094% | 38.79 |
| although/CONTRAST | 0.0211% | 0.0134% | 33.96 |
| so/CAUSE | 0.0206% | 0.0311% | 32.89 |
| while/CONTRAST | 0.0433% | 0.0331% | 28.84 |
| when/SYNCHRONY | 0.0766% | 0.0955% | 27.92 |
| also/CONJUNCTION | 0.2088% | 0.2398% | 24.35 |
| as/SYNCHRONY | 0.0564% | 0.0474% | 18.22 |
| although/CONCESSION | 0.0216% | 0.0160% | 17.52 |
| but/CONTRAST | 0.0760% | 0.0902% | 15.12 |

3.3. *Effect of text complexity on the distribution of DMs over DRs*

Once we determined that there is a difference in how DMs are used to signal a DR across corpora, we tried to verify if the distribution of DMs to signal different DRs is different across readability levels. For example, as shown in Figure 1, the DM *while* can be used in the Simple English Wikipedia to signal two DRs: CONTRAST and SYNCHRONOUS. The following examples show sentences where the DM *while* signals a CONTRAST (sentence 1); whereas in sentence 2, it signals a SYNCHRONOUS relation.

1. While [any form of energy may be conserved], [electricity is the type most commonly referred to in connection with conservation.]/CONTRAST
2. [He began his career in primary education] while [an undergraduate teaching at the Children's Community School]/SYNCHRONOUS

In the regular version of the Simple English Wikipedia corpus, each DM conveys on average 1.68 relations. On the other hand, this number decreases to 1.61 in the simple version of the same corpus. As [14] noted, in the PDTB corpus, implicit and explicit DMs combined convey on average 3.05 relations. If we only consider explicit DRs, as in our work, this number decreased to about 2.6 in the PDTB. Because of this ambiguity of DMs, we wanted to investigate how specific DMs are used to signal different DRs across different readability levels. To do so, we identified the set of relations that each DM conveys, then, the distribution of all DMs across regular and simple versions has been computed. We have used entropy in order to calculate the information of each distribution; then, used cross entropy to measure the difference between the distributions [6, 11, 5]. Formula 1 is used to calculate the entropy of the distribution of each DM (noted as H(x)) across different readability levels. Each DM is considered as a random variable, the formula. The range of values that *x* can take, noted as $r_i$ in Formula 1, are the possible DRs that the DM can signal. For example, using the DM *while* of Figure 1, the DM *x* is *while*, p($r_1$) is the probability that the DM *while* is used to signal the CONTRAST relation which is 0.706 in the regular version as opposed to 0.676 in the simple version. Similarly, p($r_2$) is the probability that the DM *while* signals a SYNCHRONOUS relations.

$$H(x) = H(p) = -\sum_i p(r_i) log(p(r_i)) \quad (1)$$

| Discourse Relation | Probability in Regular version | Probability in Simple version |
|---|---|---|
| CONTRAST | 0.706 | 0.676 |
| SYNCHRONOUS | 0.291 | 0.322 |

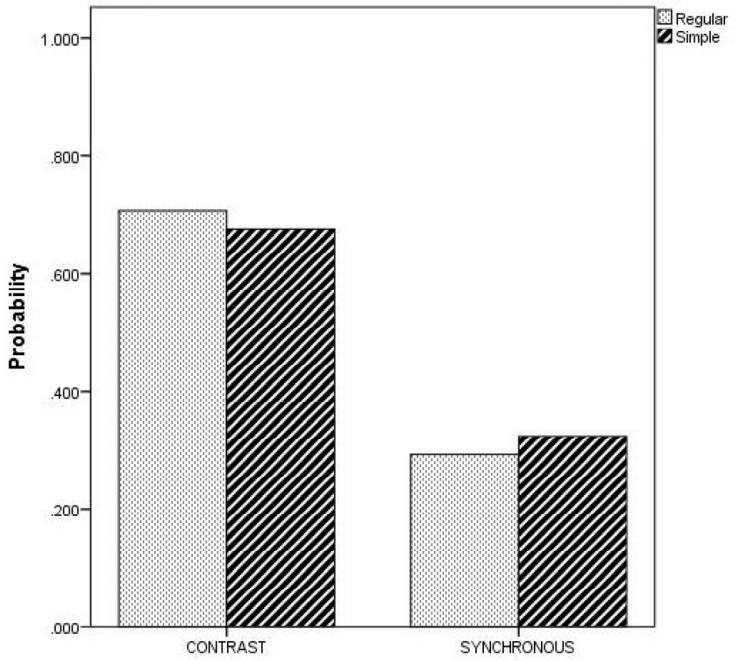


Figure 1. *Distribution of the DM while with respect to the DRs it signals across the Simple English Wikipedia corpus*

Once the entropy of each distribution has been computed, we have compared them in order to evaluate if there is a significant change in the distribution of DMs. To do so, we have used cross entropy. Formula 2 has been used for calculating the cross entropy for a specific DM called *x*. To compare two distributions using cross entropy, we assume that the first argument (*reg*) is the target probability distribution, and the other one (*simp*) is the estimated distribution that we are trying to compare against. The closer the cross entropy is to the entropy of the target distribution, the less the change in the distribution of the specific DM across readability levels. In our experiment, *reg* stands for

the regular version and $p((ri)_{reg})$ is the probability that the DM *x*, signalling the $i^{th}$ relation in the regular version; while *simp* stands for the simple version and $p((x_i))_{\text{simp}}$ is the probability that the DM *x*, signals the $i^{th}$ relation in the simple version.

$$H(reg, simp) = -\sum_i p((x_i)_{reg}) log(p((x_i))_{simp}) \quad (2)$$

The top 5 most differences in the distribution of DMs stems from the DMs in, *although, though, while* and *since*. Figure 2 shows the distribution of the DM in across the regular and the simple versions. In addition, the distribution of the DMs *although* and *though* both signalling CONCESSION and CONTRAST DRs are shown in Figures 3 and 4 respectively. As the figures show, both DMs are more frequently used to signal a CONCESSION in the simple version and a CONTRAST in the regular version. For example, the DM *although* is used 54.4% of the time to signal a CONCESSION in simple texts as opposed to 50.0% in regular texts. However, both *although* and *though* are more frequently used to signal a contrast in the regular version than in the simpler version. Finally, Figure 5 shows the distribution of the DM *since* to signal ASYNCHRONOUS and CAUSE DRs over the corpora. As Figure 5 shows, it is more probable that this DM is used to signal a CAUSE across both versions rather than ASYNCHRONOUS; however, to signal an asynchronous relation, it is more common to use since in the simple version than in the regular version.

It is interesting to note that although discourse relations seem to be preserved across readability levels (see Section 3.1), how discourse markers are used to signal these relations seems to vary across readability levels.

| Discourse Relation | Probability in Regular version | Probability in Simple version |
|---|---|---|
| ASYNCHRONOUS | 0.178 | 0.181 |
| CAUSE | 0.004 | 0.012 |
| CONCESSION | 0.002 | 0.003 |
| CONJUNCTION | 0.500 | 0.520 |
| CONTRAST | 0.093 | 0.062 |
| INSTANTIATION | 0.051 | 0.040 |
| RESTATEMENT | 0.170 | 0.178 |

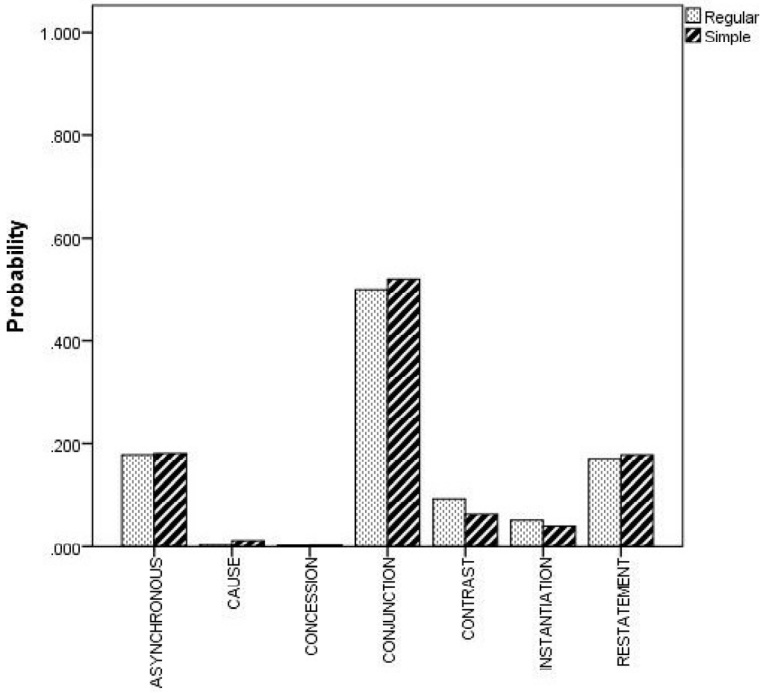


Figure 2. *Distribution of the DM in with respect to the DRs it signals across the Simple English Wikipedia corpus*

## 4. Related Work and Discussion

As [22] noted, DMs constitute valuable features to identify explicit DRs; however, they may be used in a non-discourse context. Several work have already addressed the identification, selection and placement of DMs in coherent texts (e.g. [12, 19, 8,

15, 20]). However, to our knowledge, no previous work has attempted to investigate the effect of readability level on the usage of DMs and DRs using large scale parallel corpora.

| Discourse Relation | Probability in Regular version | Probability in Simple version |
|---|---|---|
| CONCESSION | 0.506 | 0.544 |
| CONTRAST | 0.494 | 0.456 |

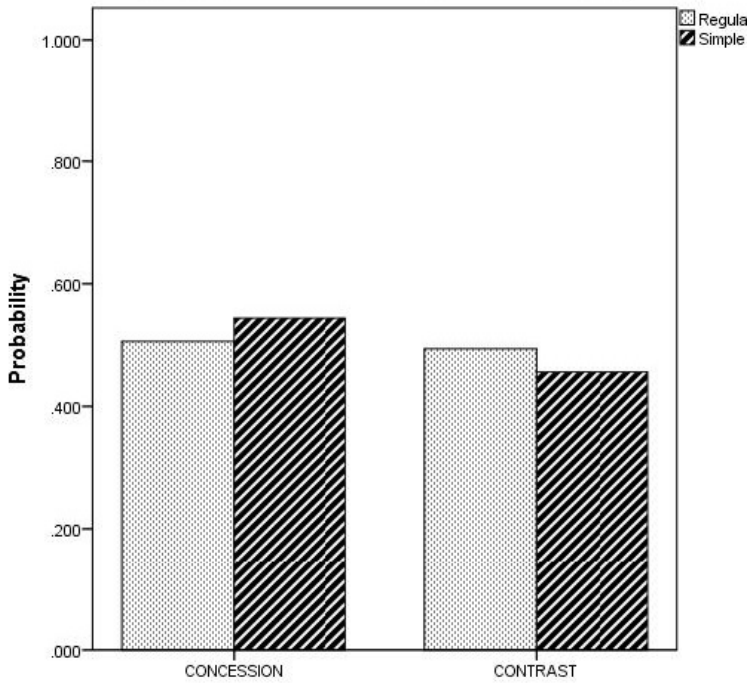


Figure 3. *Distribution of the DM although with respect to the DRs it signals across the Simple English Wikipedia corpus*

| Discourse Relation | Probability in Regular version | Probability in Simple version |
|---|---|---|
| CONCESSION | 0.394 | 0.483 |
| CONTRAST | 0.606 | 0.516 |

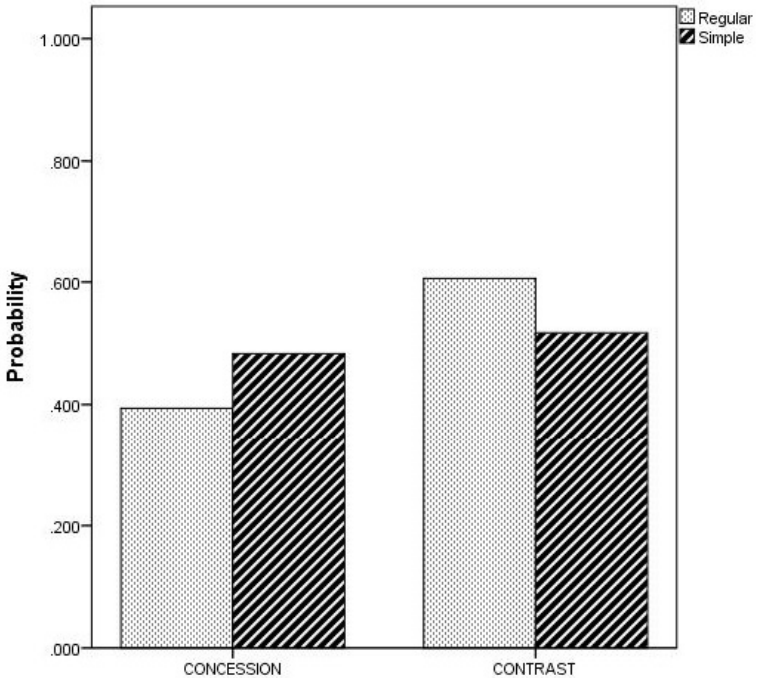


Figure 4. *Distribution of the DM though with respect to the DRs it signals across the Simple English Wikipedia corpus*

| Discourse Relation | Probability in Regular version | Probability in Simple version |
|---|---|---|
| ASYNCHRONOUS | 0.302 | 0.350 |
| CAUSE | 0.698 | 0.650 |

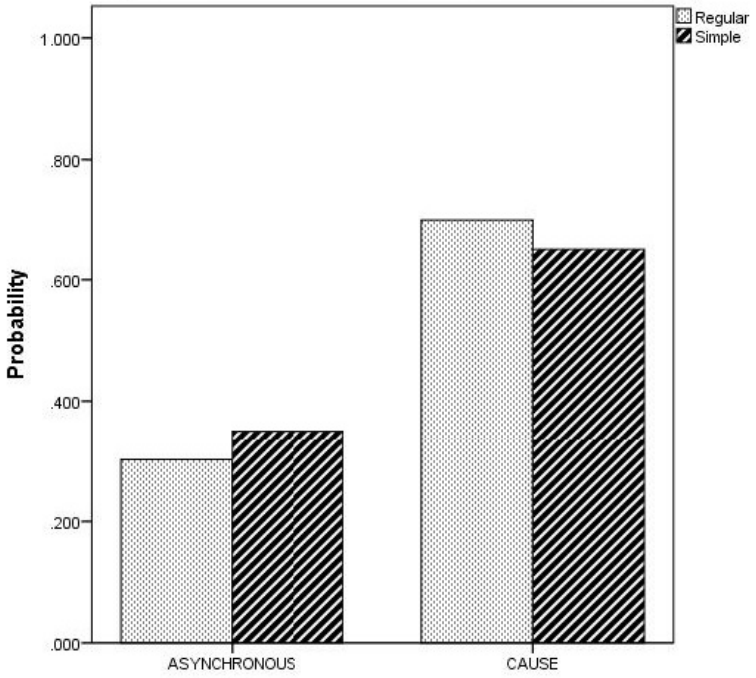


Figure 5. *Distribution of since with respect to the DRs it signals across the Simple English Wikipedia corpus*

Several attempts have been made to enhance the readability level of texts at different levels (i.e. lexical, syntactic or discourse levels) (e.g. [7, 30, 3, 24, 25]), or generating texts across different readability levels for various groups of audiences. For example, Williams' text generation system [28] generates texts at different levels of readability; however the simplification rules were based on a manual analysis of a small corpus. Three parallel texts (each with an average of 1000 to 2000 words) revealed some DMs like *so* and *but* are preferable to use in simpler texts than other DMs such as *therefore* or *hence*. She also reported that a more frequent usage of DMs result in more readable texts. This

last result seems to contradict our own (see Section 3.2) which are based on a much larger corpus.

Another related work is that of Siddharthan [25] who focused on textual simplification. Although the main focus of the work was on syntactic simplifications, Siddharthan also addressed the use specific DMs in order to increase the textual cohesion of the simplified texts. Once the original sentences were simplified syntactically, he selected specific DMs in order preserve the discourse relation between the resulting conjoined clauses. To do so, he used a set of 13 DMs and associated each DM to a single DR. The actual selection of the most appropriate DM was based on [28]'s recommendations. For example, every concession relation resulted in the use of the DM *but*. Although Siddharthan's main focus was not on discourse-level choices, a number of assumptions were made. In comparison, our work is based on a statistical analysis of a much larger corpus, uses a much larger set of DMs (the list of 100 DMs from the PDTB [22]) and does not assume a one-to-one correspondence between DMs and DRs.

## 5. Conclusion and Future Work

In this paper, we have performed an analysis of the usage of discourse relations (DRs) as well as the usage and distribution of discourse markers (DMs) across different readability levels. Our analysis of the Simple English Wikipedia corpus shows that discourse relations are preserved across different readability levels. However, the usage of discourse markers is different in the regular and their simpler counterparts. In particular, we observed that the relative frequency of DMs is higher in more complex texts. Additionally, our analysis revealed that the distribution of DMs to convey specific relations is different across different readability levels. These results seem to indicate that although the same logical and semantic information is conveyed in both simple and regular versions; how they are signalled is different.

In this article, we have analysed the changes in markers and relations at the document level, but did not look at individual

changes. As future work, it would be interesting to investigate discourse relations and discourse markers across specific sentence alignments in order to analyse changes in their individual usage. For example, under which conditions, a concession is changed to a condition at different readability levels.

## REFERENCES


1. Bachand, F.H., Davoodi, E., Kosseim, L. 2014. An investigation on the inuence of genres and textual organisation on the use of discourse relations. In proceeding of the *15th International Conference of Computational Linguistics and Intelligent Text Processing (CICLing), LNCS-volume 8404*, (pp. 454-468). Springer.
2. Carlson, L., Okurowski, M.E., Marcu, D. 2002. RST discourse treebank. *Linguistic Data Consortium*, Catalog Number-LDC2002T07.
3. Chandrasekar, R. & Srinivas, B. 1997. Automatic induction of rules for text simplication. *Knowledge-Based Systems*, 10/3, 183-190.
4. Coster, W. & Kauchak, D. 2011. Simple English Wikipedia: A new text simplification task. In proceedings of the *49th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies (ACL-HLT)* (pp. 665-669), Short papers-Volume 2. Portland, Oregon, June 2011)
5. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. John Wiley & Sons.
6. De Boer, P. T., Kroese, D. P., Mannor, S. & Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134/1, 19-67.
7. Devlin, S. & Tait, J. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases* (pp. 161-173).
8. Di Eugenio, B., Moore, J. D. & Paolucci, M. 1997. Learning features that predict cue usage. In proceedings of *EACL* (pp. 80-87), Madrid, Spain.
9. Feng, V. W. & Hirst, G. 2012. Text-level discourse parsing with rich linguistic features. In proceedings of the *50th Annual Meeting of the Association for Computational Linguistics*: *Long Papers - Volume 1* (pp. 60-68), ACL-2012.

10. Hernault, H., Prendinger, H., A. duVerle, D. & Ishizuka, M. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1/3.
11. Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical Review*, 106/4, 620.
12. Knott, A. 1996. A data-driven methodology for motivating a set of coherence relations. PhD dissertation, University of Edinburgh.
13. Laali, M., Davoodi, E. & Kosseim, L. 2015. The CLaC discourse parser at CoNLL-2015. In N. Xue, H. T. Ng, S. Pradhan, R. Prasad, C. Bryant, A. Rutherford, (Eds.), *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015* (pp. 56-60), Beijing, China, Jul 30-31.
14. Laali, M. & Kosseim, L. 2014. Inducing discourse connectives from parallel texts. In the *25th International Conference on Computational Linguistics (COLING)* (pp. 610-619), Dublin, Ireland.
15. Lin, Z., Kan, M. Y. & Ng, H. T. 2009. Recognizing implicit discourse relations in the Penn discourse treebank. In proceedings of the *2009 Conference on Empirical Methods in Natural Language Processing* (pp. 343-351), Volume 1, Singapore.
16. Lin, Z., Ng, H. T. & Kan, M. Y. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20/2, 151-184.
17. Mann, W. C. & Thompson, S. A. 1987. Rhetorical structure theory: A framework for the analysis of texts. Tech. rep., *IPRA Papers in Pragmatics*, 1.
18. Meyer, T. & Webber, B. 2013. Implicitation of discourse connectives in (machine) translation. In proceedings of the *1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)* (pp. 19-26), Sofia, Bulgaria.
19. Moser, M. & Moore, J. D. 1995. Investigating cue selection and placement in tutorial discourse. In proceedings of the *33rd annual meeting on Association for Computational Linguistics* (pp. 130-135), Cambridge, Massachusetts, USA.
20. Patterson, G. & Kehler, A. 2013. Predicting the presence of discourse connectives. In EMNLP (pp. 914-923), Seattle, Washington, USA.
21. Pitler, E. & Nenkova, A. 2009. Using syntax to disambiguate explicit discourse connectives in text. In proceedings of the *ACL-IJCNLP 2009 Conference Short Papers* (pp. 13-16), Suntec, Singapore.

22. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. L. 2008. The Penn discourse treebank 2.0. In proceedings of the *6th International Conference on Language Resources and Evaluation (LREC)*, Marrakesh, Morocco, Jun.
23. Rayson, P. & Garside, R. 2000. Comparing corpora using frequency profiling. In proceedings of the *Workshop on Comparing Corpora* (pp. 1-6), Hong Kong, Oct.
24. Scarton, C., De Oliveira, M., Candido Jr, A., Gasperin, C., Aluisio, S. M. 2010. Simplifica: A tool for authoring simplified texts in brazilian portuguese guided by readability assessments. In proceedings of *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Demonstrations* (pp. 41-44), Los Angeles, CA, USA, Jun.
25. Siddharthan, A. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4/1, 77-109.
26. Soricut, R. & Marcu, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In proceedings of the *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)* (pp. 149-156), Vol. 1., Edmonton, Canada, Jun.
27. Webber, B. 2009. Genre distinctions for discourse in the Penn TreeBank. In proceedings of the *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-AFNLP)* (pp. 674-682), Vol. 2., Suntec, Singapore, Aug.
28. Williams, S., Reiter, E. & Osman, L. 2003. Experiments with discourse-level choices and readability. In proceedings of the *9th European Workshop on Natural Language Generation (ENLG-2003)* (pp. 127-134), Budapest, Hungary, Apr.
29. Yamangil, E. & Nelken, R. 2008. Mining wikipedia revision histories for improving sentence compression. In proceedings of the *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies (ACL-HTL): Short papers* (pp. 137-140), Columbus, Ohio, Jun.
30. Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C. & Lee, L. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 365-368), Los Angeles, California, USA, Jun.

**ELNAZ DAVOODI**
DEPARTMENT OF COMPUTER SCIENCE
AND SOFTWARE ENGINEERING,
CONCORDIA UNIVERSITY, MONTREAL, CANADA.
E-MAIL: <E_DAVOO @ENCS.CONCORDIA.CA>

**LEILA KOSSEIM**
DEPARTMENT OF COMPUTER SCIENCE
AND SOFTWARE ENGINEERING,
CONCORDIA UNIVERSITY, MONTREAL, CANADA.
E-MAIL: <KOSSEIMG@ENCS.CONCORDIA.CA>